Trustworthy Bias Measures for Language Models and Word Embeddings Natural Language Processing 2

Juno Prent
Universiteit van Amsterdam
11915307@uva.nl

Wu Wang Yang
Universiteit van Amsterdam
ID:14269244

Wenkai Pan
Universiteit van Amsterdam
wenkai.pan@student.uva.nl

Abstract

This study examines the reliability and validity of the Word Embedding Association Test (WEAT) as a measure of gender bias, applying a selection of experiments on it that have not been previously reported on. We employ pretrained word2vec embeddings in our experiments, as WEAT is specifically designed for static embeddings. The findings indicate that WEAT's measurements consistently detect a similar gender bias present in these embeddings. Moreover, the results demonstrate consistency with several other bias measures, showing WEAT to be reliable and valid.

1 Introduction

In everyday language, many biases of different types occur, such as gender, race or religion. This can lead to a perpetuation of incorrect stereotypes or other misinformation, among other negative effects. Unfortunately, these language biases are also captured by natural language processing (NLP) models and data.

Here, the focus will lie on the detection of gender bias in static word embeddings from word2vec[6]. To do so, the WEAT[2] bias metric is used, as implemented from the WEFE[1] Python package. These bias scores will be analyzed on two different aspects. These are reliability, so whether the measurements are consistently similar, and validity, which determines whether a certain bias score for embeddings means that these are actually biased, or whether this score might actually be influenced by other factors.

The analysis of WEAT's bias measurements leads to the following research questions:

1. What is the reliability of WEAT's bias measurements?
2. What is the validity of WEAT's bias measurements?

These were answered through experiments pertaining to WEAT's reliability and validity, assessing how consistent its measurements are and how they compare to those of different bias metrics. From these experiments, it was found that WEAT is indeed a reliable and valid bias metric.

2 Background

In recent years, the issue of bias in language models and word embeddings has received significant attention. Several studies have highlighted the presence of gender biases in these models and their potential negative implications for perpetuating stereotypes and misinformation. Gonen and Goldberg [5] conducted a comprehensive analysis of debiasing methods. They demonstrated that existing debiasing techniques may mitigate explicit gender bias in word embeddings, but fail to address the underlying systematic biases. This highlights the need for effective and reliable bias measurement methods.

Evaluation of gender bias in word embeddings is a crucial aspect of addressing bias in language models. However, Orgad and Belinkov [7] identified flaws in gender bias evaluation methods. They discussed the limitations of the popular WEAT and proposed alternative evaluation approaches to provide a more nuanced understanding of gender bias in word embeddings. This highlights the importance of critically examining bias measurement methods and exploring alternative metrics.

Ethayarajh, Duvenaud, and Hirst [4] delved into the issue of undesirable word embedding associations. They highlighted the presence of unintended associations between protected attributes (e.g., gender, race) and other words in word embeddings. Their work emphasized the need to develop measures that not only capture explicit biases but also account for implicit associations and undesirable connotations present in word embeddings.

Guarding against biases in protected attributes

[1] https://pypi.org/project/wefe/

is another significant concern. Ravfogel et al. [8] proposed the Iterative Nullspace Projection method. Their approach focused on protecting sensitive attributes, such as gender or race, by iteratively projecting the embeddings onto a null space to remove associations with the protected attribute. This work highlights a potential technique to mitigate biases associated with protected attributes in word embeddings.

Drawing from the insights of these studies, this research aims to address the limitations of existing bias measurement methods and explore the reliability and validity of WEAT as a measure of gender bias. Additionally, we aim to investigate the consistency of WEAT's measurements and compare them with other bias metrics. By doing so, we aim to provide a deeper understanding of gender biases present in word embeddings and contribute to the development of more effective and trustworthy bias measurement techniques.

3 Approach

3.1 WEAT

WEAT is a bias metric aimed at detecting bias within static word embeddings. As its input, it takes two different types of sets: attribute sets, which can be gendered words, such as pronouns or names, and target sets, which are supposed to be gender-neutral words, such as occupational terms.

To obtain its output, WEAT then calculates the cosine similarity between the attribute and target sets. A higher value for the cosine similarity can be interpreted as a higher level of bias, as this means there is more of an association between the embeddings of the gender-neutral and gendered terms. This, of course, is often not desirable.

3.2 Assessment methods

3.2.1 Reliability

To test the reliability, the choice was made to use the method of internal consistency. Internal consistency is a measure used to assess the consistency and reliability of a test or measurement instrument, such as WEAT. It examines the extent to which the items or components within the test are measuring the same underlying construct or concept. For these experiments, random samples of various sizes of the target and attribute concept lists were taken and their effect sizes compared to an established baseline.

Employing this internal consistency reliability measure gives insight into whether WEAT is consistently measuring the desired associations between target and attribute concepts, namely gender bias. Ideally, the effect size for randomized samples is still close to that of the baseline, given a sufficiently large sample size. If this isn't the case, WEAT might be overly dependant on specific items to obtain its results, leading to very different results when these items are not present. Higher levels of internal consistency provide evidence for the reliability of the metric.

3.2.2 Validity

To test the validity of WEAT, we chose to employ convergent validity. Convergent validity is a type of validity assessment that examines the degree to which a test or measurement correlates with other measures or indicators that are expected to assess a similar concept, in this case bias. In the context of testing the validity of WEAT, we employ convergent validity to evaluate whether the results of WEAT align with other measures that assess similar associations between target and attribute concepts. For this, the measures Word Analogy Testing and Word Similarity Comparison were used.

Word Analogy Testing is a method used to evaluate the performance of word embeddings or language models in capturing semantic relationships between words. It involves assessing the ability of a model to accurately complete analogical relationships in the form of word analogies and can also be used to assess gender bias in word embeddings or language models. To evaluate the gender bias, we construct analogy questions that specifically target gender-related associations. An example would be "father is to son as mother is to __".

Word Similarity Comparison is the process of quantifying the similarity between two words based on their semantic or contextual similarity. It involves measuring how closely related or similar two words are in terms of gender association.

To support this experiment, a downstream task was conducted: semantic textual similarity for bias (STS-B)[9], a common benchmark for assessing how NLP models measure the semantic similarity between pairs of texts. Here, a modified version was used, as used in the assignment notebook. It compares similarity scores between three pieces of text. For example, if we have two differently gendered sentences A and B, along with a neutral sentence C, similarity scores of A and C with B and

C can be compared. The gender bias of the neutral term in C is then the difference in the similarity averaged over a set of template sentences. Figure 1 further illustrates this approach.

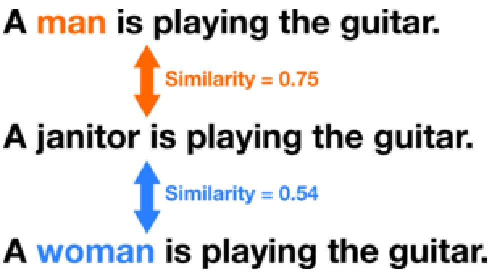
A man is playing the guitar.
Similarity = 0.75
A janitor is playing the guitar.
Similarity = 0.54
A woman is playing the guitar.

Figure 1: Intuition behind STS-B

3.3 Word embeddings

For this undertaking, the decision was made to utilize word2vec embeddings that were pretrained specifically on the Google News corpus. This particular model, known as *word2vec-google-news-300*, is a pretrained word embedding model developed by Google and trained on an extensive collection of Google News articles. It encompasses word vectors with a dimensionality of 300. To access this pretrained word2vec modem, *downloader* from the *Gensim* library was used.

4 Experiments and results

4.1 WEAT baseline

To obtain a baseline for WEAT that could be used for further comparisons, WEAT was applied to the word2vec embeddings. Four lists of gendered words and associated concepts were created, each with 32 items. Their categories were male names, female names, family relationships, and occupations. For the names, the top 32 most common names per gender in the past century in the USA[2] were chosen. For simplicity, WEAT is conducted using the aforementioned *WEFE* package, which is developed for the evaluation and analysis of fairness in word embeddings. It provides a range of tools and metrics to measure and quantify different aspects of bias and fairness in word embeddings. The reason why we chose WEFE is because it comes with handy functions to assess various types of biases, such as gender bias, racial bias, and other forms of social bias, in word embeddings.

WEAT yielded an effect size of approximately 1.22. The effect size quantifies the magnitude of the association between the target and attribute concepts in WEAT. A larger effect size suggests a stronger association. In this case, the effect size implies that the difference between the means of the target word embeddings in the two attribute categories is well over one standard deviation. This suggests a notable distinction between the two sets and indicates a substantial effect or relationship, that being a relatively strong gender bias.

4.2 Reliability

To measure the reliability of the WEAT measurement, an internal consistency experiment was done in which randomized samples of varying sizes, ranging from 4 to 24, were taken of each of the four concept lists. This was done 10.000 times for each sample size, giving a total of 210.000 sets of four randomized samples.

WEAT was then applied on these samples each time to determine its effect size and how close it was to the previously established baseline. The average effect size obtained for each sample size, as well as the baseline's effect size, is visualized below in Figure 2.

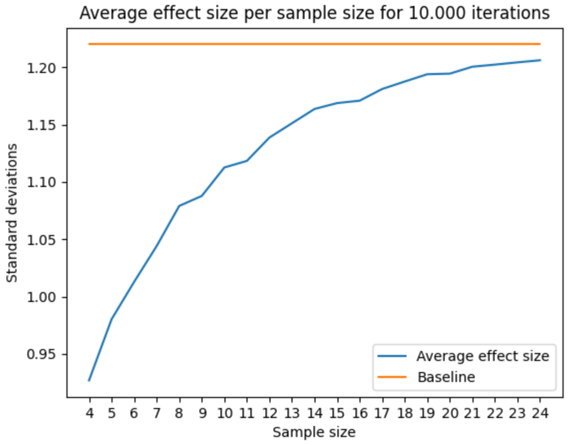

Figure 2

4.3 Validity

4.3.1 Word Analogy Testing

For Word Analogy Testing, we defined a list of 10 analogies ourselves which can be found in the notebook. The results were obtained as follows: if the model predicts an analogy as expected then the result is recorded as the index of this sample. Otherwise, it is recorded as 0. For example, if the data set has 4 samples, in the case where the first three samples are correctly predicted and the

[2] https://www.ssa.gov/oact/babynames/decades/century.html

last sample is predicted wrong, the results will be [1,2,3,0].

4.3.2 Word Similarity Comparison

For Word Similarity Comparison, a list of gendered pairs and gender-neutral pairs are defined, each containing 78 pairs. Results are also recorded in the same way as that in Word Analogy Testing.

4.3.3 Evaluation

To analyze the outcomes of the aforementioned tests, we utilized the Kappa score as a means to assess the level of agreement between the two evaluations. This is a statistical measure employed to gauge the agreement between two raters or evaluators categorizing items into distinct categories. The Kappa score goes from 0 to 1, with 1 signifying complete agreement.

Here, the computation of the Kappa score is done using the *cohen_kappa_score* library in *sklearn* package. With this, a score of approximately 0.76 was obtained. This falls in the 0.61-0.80 segment, which indicates that there is a substantial level of agreement.

4.3.4 STS-B

For this undertaking, we utilized a pre-existing dataset comprising a total of 276 samples[3]. Each sample consists of three sentences that incorporate either the term "man," "woman," or an occupational descriptor, as described previously. The results show that 69.9% of the input samples favour the male concept.

5 Discussion

5.1 Reliability

When looking at the results of the internal consistency experiment seen in Figure 2, the consistency of WEAT's measurements becomes apparent immediately. Keeping in mind the baseline's effect size of around 1.22 standard deviations, it can be seen that this value is quickly approached as the sample size increases, with only the very lowest sizes being far off the baseline. When the sample size is 16, meaning exactly half of each concept list is randomly sampled each iteration, the difference with the baseline is less than 0.05 already.

From this, it can be concluded that WEAT's measurements are indeed reliable. There is no overdependence on specific input elements and measurements are consistent with the baseline over large amounts of iterations, for differing sample sizes. As a bias metric should be reliable to be of any importance, this does not come as a surprise.

5.2 Validity

The approximately 0.76 Kappa score from the convergent validity experiment suggests that the two additional measures substantially align with each other on the assessment that the pretrained word2vec embeddings possess gender bias. This also aligns with the outcomes of WEAT's measurement of the same embeddings.

The outcome of the downstream STS-B task further solidifies these findings. In any unbiased setting, there would not be such a stark preference for male concepts in words that are supposed to be gender-neutral in meaning. This should thus have lead to only a theoretical 50% preference for the male concepts, which is far lower than was obtained. As this indication of gender bias is also in line with WEAT, it can be concluded that there are many signs that WEAT's bias measurements are valid. As this is desired for a bias metric such as WEAT, this does not come as a surprise either.

5.3 Future work

As the focus on bias detection is a recent development, it is not quite clear which bias metrics can be fully trusted yet[1]. As such, future work might do well to focus on establishing whether bias metrics other than WEAT, such as Embedding Coherence Test[3], can also be considered reliable and valid. Doing so for a wider selection of bias metrics will result in a clearer overview of trustworthy methods for detecting biases for a wider array of use cases.

6 Conclusion

To summarize, several experiments were performed with the bias metric WEAT to assess it on two aspects, those being the reliability and validity of its gender bias measurements. From the results, it became clear that these measurements are consistent among different randomized settings and also in line with the bias measurements of several other bias metrics. As such, it can be concluded that WEAT is indeed a reliable and valid metric for gender bias.

[3] https://drive.google.com/drive/folders/1PQlC1P1zhhgtGc7dSD5UcIB9IZtEzPi7

References

[1] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[3] Sunipa Dev and J. M. Phillips. Attenuating bias in word vectors. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[4] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[7] Hadas Orgad and Yonatan Belinkov. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington, July 2022. Association for Computational Linguistics.

[8] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics.

[9] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032, 2020.

Appendix

Contribution overview

Juno:

- Did the tutorial notebook, except for Q1.3. Wrote the PDF file about it
- Wrote reliability experiment
- Implemented STS-B data set
- Wrote Introduction, all sections relating to reliability experiment, Discussion and Conclusion
- Wrote other small parts, restructured/rewrote large parts of our original report to adhere to rubric structure

Wangyang:

- Did the tutorial Q1.3
- implementation for WEAT for pretrained word2vec embeddings
- Designing and coding for validity experiment
- Wrote Abstract, WEAT and word embeddings sections in Approach, all sections relating to validity experiment

Wenkai:

- Wrote Background section